

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

**CONCEPTUAL MODELLING OF ADMINISTRATIVE REGISTER INFORMATION AND
XML - TAXATION METADATA AS AN EXAMPLE**

Invited Paper

Submitted by Statistics Finland¹

I. INTRODUCTION

1. As register-based statistics production is becoming more general, one of the main factors influencing the use of register data is the content of register data. Content management of the existing register data used in statistics production has a direct effect on the end result of the use of administrative data, which in this case means statistical information, its quality and extent. Similarly, content management of register data essentially influences the key stage of statistics production, i.e. editing and its end result.

2. The first issue we address is the question of how to form a conception of the information content of the administrative register. At present the means for this are fairly limited and quite conventional in their use. The usual way is to look up a description of administrative data from fairly short administrative register descriptions produced by authorities themselves and from printed handbooks or their electronic copies. For statistics production these descriptions and administrative handbooks are external and separate information sources that are difficult to handle in information technology and are managed manually.

3. In addition to being about an adequate informative basis for editing, it is also a question of whether contentual information created in editing can be saved as descriptions of the results of editing work to be used further in later stages of statistics production. It is not only a question of that producers of statistics have full contentual command of the register data but also how the content of the statistical information collected from register sources is described and transmitted to users of statistics so that they can form a correct perception of the content of statistical information and have a possibility to interpret and use the thus formed statistical information as accurately as possible.

4. It is a process where a conception of the register data must be passed on to statistics producers and they must be able to transfer that information to users of statistics in a form understandable to them. The only way to transmit the meaning and purpose of the data is to ensure easy availability of adequate content information in all stages of data processing. The challenge is how the present process, where the description of administrative data can mostly be read from the authorities' administrative handbooks (see Figure 1.), can be transformed into such that it meets the requirements for the usability and presence of

¹ Prepared by Heikki Rouhuvirta, Statistical Methodology R&D, Statistics Finland; heikki.rouhuvirta@stat.fi

the contentual description of data both in the production process to statistics producers and in the distribution of statistical information to users of statistics.

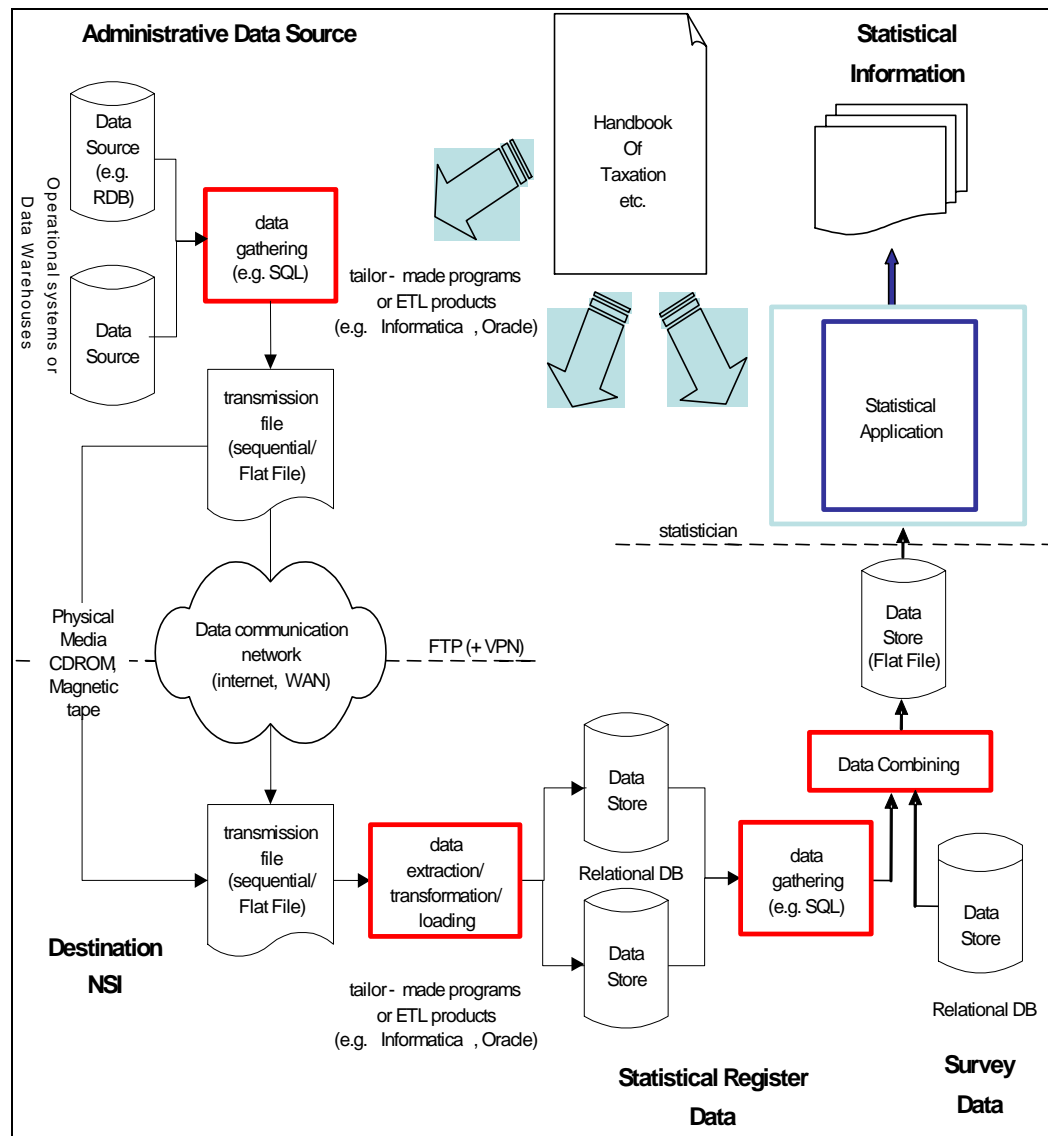


Figure 1. Present state of compilation of administrative data

5. In case these external information sources for content and the new contentual information produced in editing can be flexibly integrated as part of technological information systems of statistics production, such integration could at the same time intensify statistics production as well as improve its quality.

II. DATA SEMANTICS OF ADMINISTRATIVE REGISTER

6. In order to better integrate the contentual information into statistics production in its different stages, we need to extend the conventionally used data models so that they will also include a semantic description of the data. Similarly, the data model of the register information used in data processing should contain a semantic description of administrative data as fully as possible. Secondly, data models should be revised so that they would allow optimally unambiguous and contentually permanent transferability for the semantic definition. Thirdly, the revised data models should enable presentation of

statistical information and register data semantics in a uniform reference frame and in as standard manner as possible.

7. The methods and technologies conventionally used in statistics production did not as such point at any clear solution model for the intended extension and editing of the data model for administrative register information. The situation was also the same for the technologies used by producers of administrative data in their activity. They could not offer any ready-made solutions that could be applied as semantic data models for administrative register information.

8. In absence of ready-made applicable models it was quite natural to look for solutions from the starting points applied at Statistics Finland for conceptualisation of statistical metadata². This is quite reasonable also because in semantic data it is partly a question of metadata, or to put it more accurately, of metadata in respects where the metadata is semantically significant. In addition, this application model offers a fundamental advantage as the concrete solution of statistical metadata is the one into which the semantic description of administrative information should be integrated.

III. CoSSI AS THE REFERENCE FRAME OF METADATA

9. In modelling of statistical information³ the starting point of the definition of metadata is that in the conceptualisation of the contentual description of statistical information use is made as far as possible of the concepts characteristic of statistical information, the concepts and concept structures it contains and the logic that allows sufficiently multifaceted and complex concept structures for an exhaustive description of the information content. As the used description method allows implementation of complicated structure descriptions, the procedure does not have essentially any factors that would per se somehow force to contract or limit the contentual description.

10. Correspondingly, attention was paid in the conceptualisation of information content of administrative material to the simplification of the characteristic concepts of the administrative information in question and to the definition of the logical structure of the concepts and the concept model. By analysing the information related to administrative data and the existing contentual information about it to the extent that it is produced and accumulated by administration, the aim was to produce a fully defined and logical information structure characteristic of the current administrative information, within the frame of which the semantic description of administrative data can be produced.

11. The first approach was applied to the personal taxation material and to the administrative information describing it, from which the results presented here are taken⁴.

IV. TAXATION AS AN INFORMATION SYSTEM

12. Taxation is generally seen as a process of tax collection, which results in paid taxes. Taxation as a process includes several different types of information recorded in taxation, such as that describing reporting of income, determination of taxes, complaints about them, etc. This standpoint is not very useful in this connection because the process view of taxation obscures taxation as a conceptual information system.

13. As we are not interested in the collected taxes as such, i.e. taxes paid, but in what kind of information taxation as a process is based on, we can get closer to the content of tax information by viewing taxation from the determination bases of taxes, forgetting now the information related to the implementation of taxation and the information saved on that. The latter would naturally be interesting

² See Rouhuvirta H., An alternative approach to metadata – CoSSI and modelling of metadata, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636.

³ See for further details Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.9, Statistics Finland 2003.

⁴ Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S., Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Codacmos 2004. Project IST-2001-38636.

only in case we intended to describe, explore or study the case-specific taxation practice and its problematics. But our purpose lies elsewhere: to produce a description of the data contained in the taxation material that will universally explain what is the actual meaning of the data collected to the register on payment of taxation.

14. In addition to the determination basis of taxes, i.e. the income concept, the logical basic components of taxation include deduction and the tax itself as the end result of calculation. When tax is not examined, we end up with the concept space of income, or rather that formed of income sources and deductions presented in Figures 2 and 3.

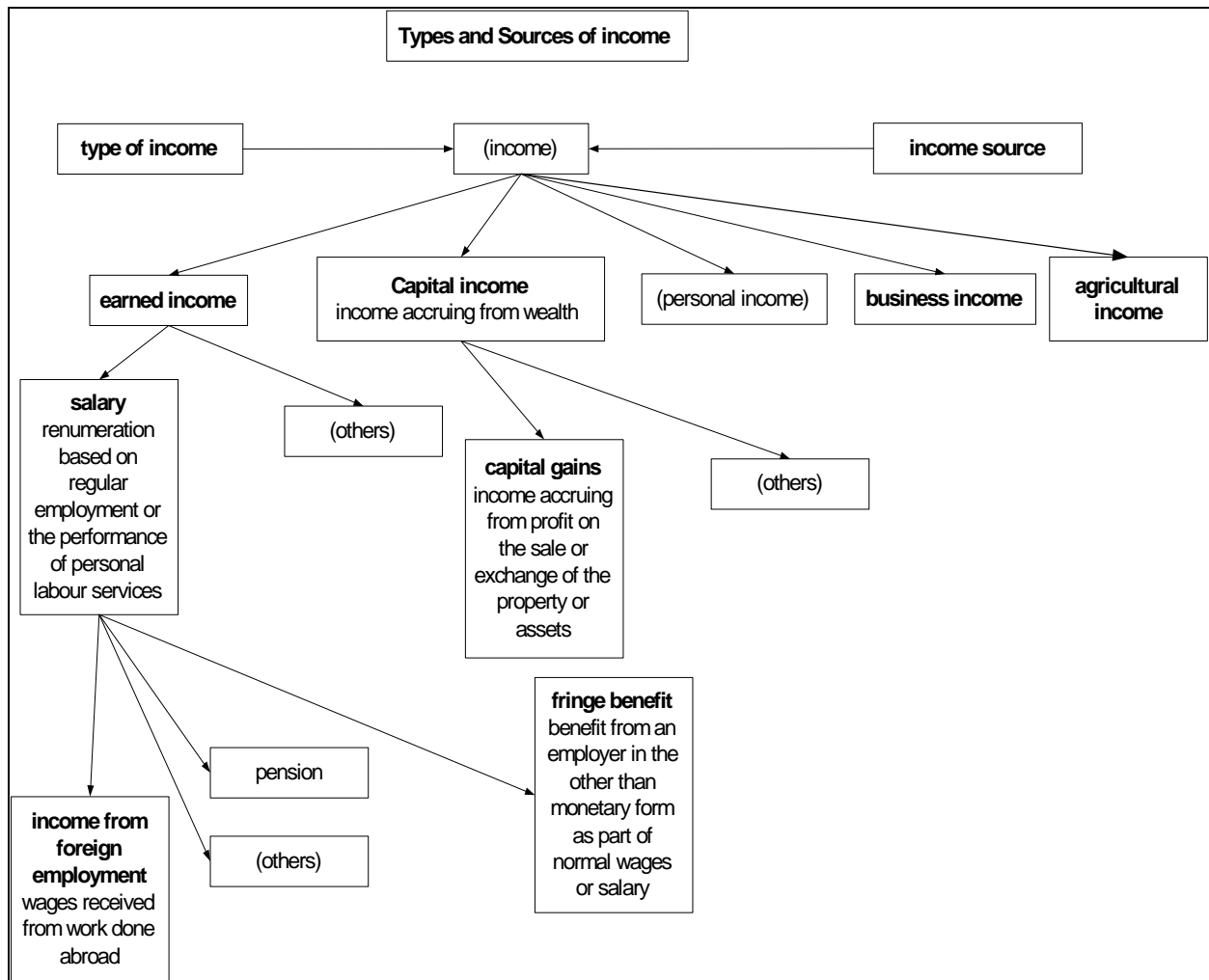


Figure 2. Types and sources of income

15. The types of income are earned income and capital income. Income source indicates the activity in which the income of a taxpayer originates. A taxpayer may have three sources of income: business (business income), agriculture (agricultural income) or other activities (personal income). The source of income determines which law is applied in calculating a taxpayer's taxable income.

16. Deductions can be made either from income or taxes. Deductions made from income include such as deduction for pension insurance premiums, discretionary allowance for circumstantial incapacity to pay taxes, pension income allowance, earned income allowance and low-income allowance. The deductions made from the taxable income are, for example: child maintenance credit, domestic help credit and credit for capital income deficit.

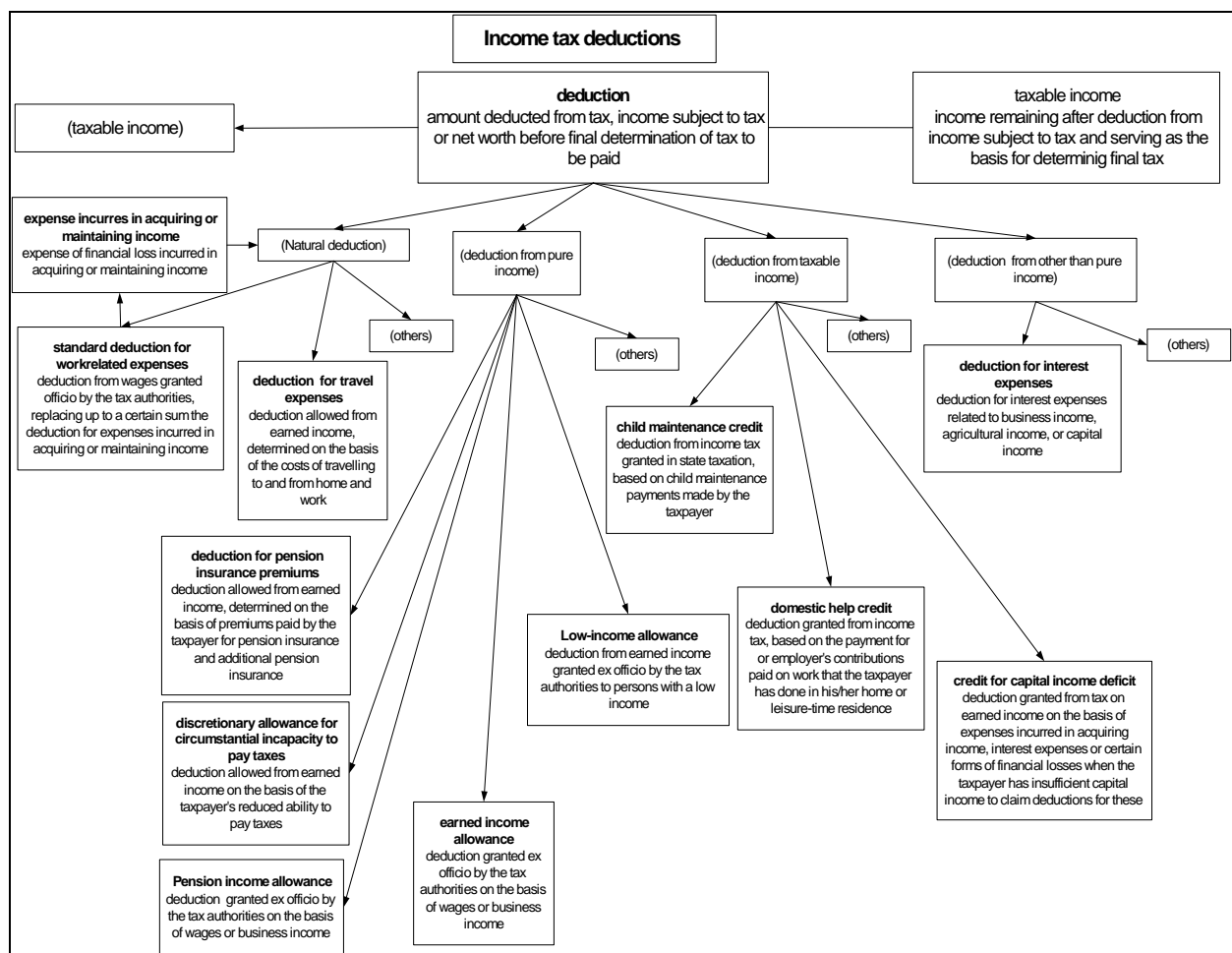


Figure 3. Income tax deductions

V. SOURCES OF CONTENT INFORMATION FOR TAXATION

17. Although all the data handled in taxation is collected with various tax forms, the information defining the data collected on the forms is contentually fairly concise. The main contentual information concerning taxation is stored in the Tax Administration's various instructions and handbooks. The Tax Administration also has technical descriptions of its information systems. The joint reference frame for all these information sources is formed by the taxation legislation. In Finland legislative texts are stored in a public electronic database on acts and decrees accessible to all through the Internet.

18. The contentually largest source for the information describing personal taxation is the Tax Administration's handbook on personal taxation. The handbook is a printed publication and it is used at Statistics Finland, for instance. The size of the handbook is about 800 pages. It describes in detail the entire personal taxation practice and the related legal cases. The handbook also has references to the legislation concerning personal taxation. An updated version of the handbook is produced separately for each starting tax year.

19. There is an electronic original of the handbook mainly for printing, which is also used for storing the information to be updated. For production of the electronic original there is a document definition, which mainly describes the external structure of the handbook. A printed publication can thus be produced automatically from the original. The scarce contentual structure in the document definition (e.g. legal cases and search words) does not however follow the contentual logic of taxation information.

VI. FROM SEMANTIC DATA TO STRUCTURED METADATA

20. In addition to the theoretical bases certain practical matters also have an effect on editing of the concept model for semantic taxation information. The concept structure to be created for taxation information is formed into such that in its scope it is easy and widely possible to utilise the existing electronic data defining the information content and to transfer this information to the user of taxation data without losing the information. In addition, the idea was to employ the structure definition so that it is in a technologically suitable form for information processing.

21. From the presented starting points a logical concept model for taxation information was defined, which as a tree structure⁵ is shown in the figures below (see Figures 4a. and 4b.).

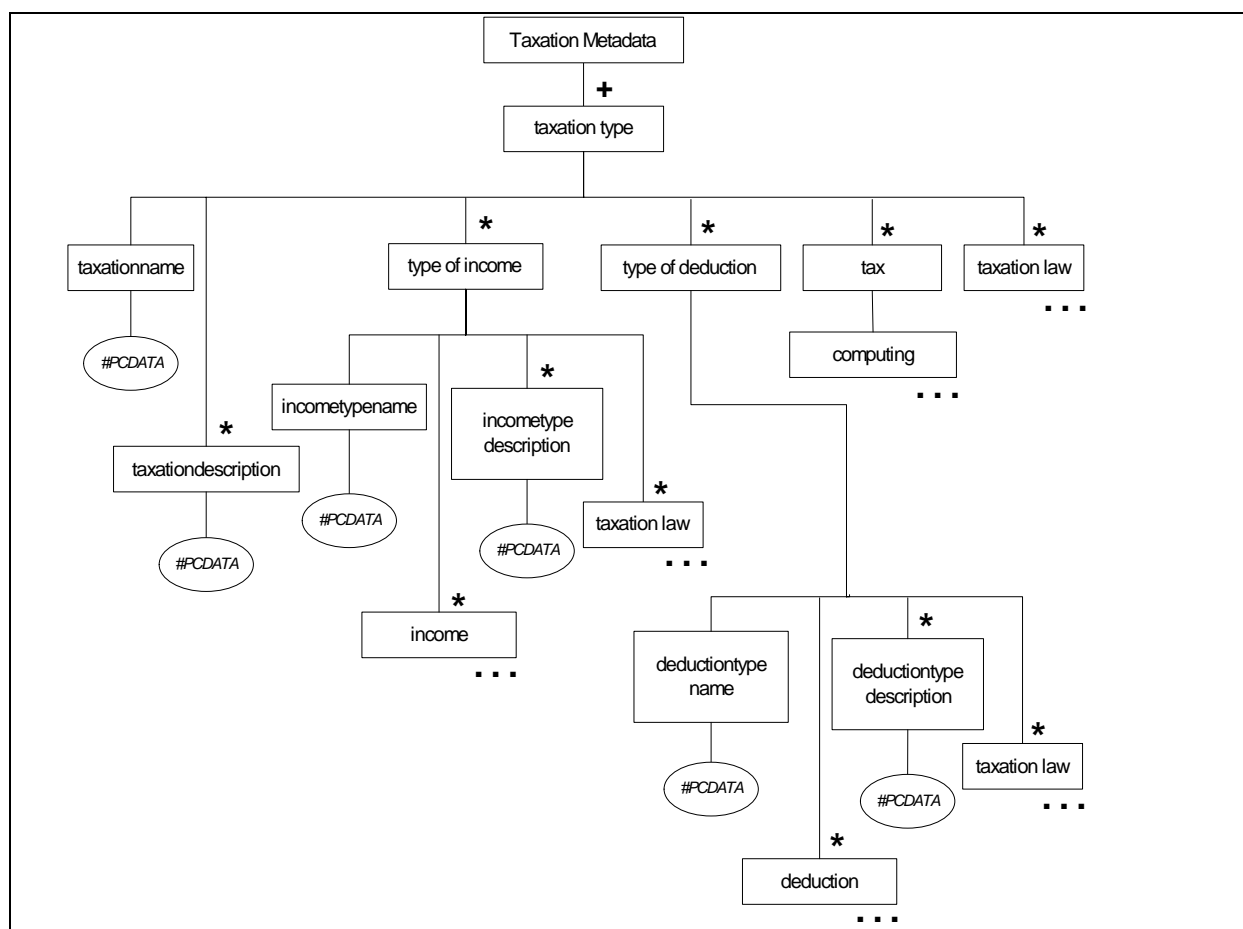


Figure 4a. Logical concept model for taxation

22. Within the framework of the structure it is possible to describe the key concepts of taxation except for the process description of taxation, which was excluded from the analysis as a factor defining the semantic content of information secondarily and thus omitted from the concept model. The detailed structure definition was implemented as XML DTD⁶. The XML definition was also used as the basis for application development. In its present form, the implemented detailed structure definition does not

⁵ Method of tree structure description, see Maler, E. & El Andaloussi, J. 1996. Developing SGML DTDs. From text to model to markup. Upper Saddle River (NJ), USA: Prentice Hall.

⁶ Rouhuvirta, H. and Lehtinen, H., Taxmeta DTD. Codacmos 2004. Project IST-2001-38636.

comprise tax calculation algorithms, but they can be included in the structure definition as the concept model for determination of taxes to be paid.

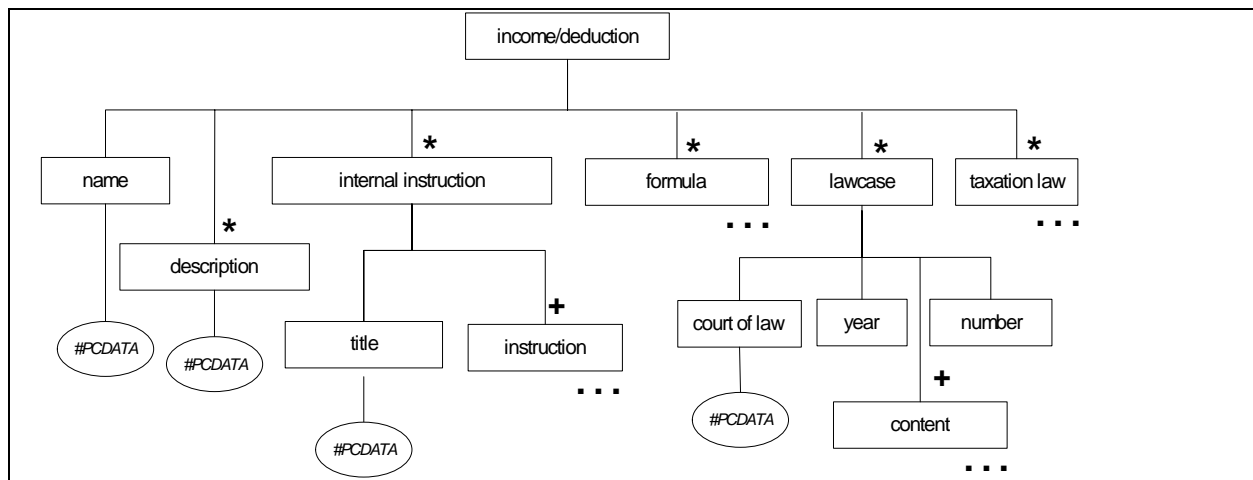


Figure 4b. Logical concept model for taxation – income and deduction

23. In the implementation the information structure of taxation is hierarchical. The hierarchy of taxation metadata is determined by tax type that contains income items and deduction items and the contentual information concerning them. The structure of the information consists of name, description and instruction. In addition, it is possible to attach to different hierarchy levels descriptions of the legal cases concerning the section in question and references to the legislation regulating it.

24. Within the scope of the structure definition it is possible to present the information contained in the taxation handbook following the contentual logic of taxation and taxation information. Information contents may change in the scope of the structure. For example, formation of new taxation items as the legislation changes yearly or production of new taxation information when the tax authorities review their instructions do not yet compel to make any changes to the structure of the information model. The need for changing the structure arises only when a completely different philosophy and the relevant logic are adopted in taxation.

VII. STRUCTURED METADATA - REGISTERS

25. The electronic version of the personal taxation handbook was used in testing the information model for register information. The electronic information content was arranged on a one-off basis according to the structured information model. The use of the information was tested in the development environment⁷, where structured register metadata was linked to the information selected from the relational database (see Figure 5. and 6.).

26. The first of these user interface images (see Figure 5.) shows the information of one person's tax register record in the relational database: first is given the code used in the register, second the value in euro and the third field has a plain-language register code, which can be either a column code of the relational table or an individualising code for the calculated information content. The information identified by a code in the relational database is such as pay, proceeds from sale of timber, membership fees paid, etc. Metadata can be searched for each tax record field by using these codes.

⁷ The software solution of the development environment was made by Laavola and Harlas from Tietokarhu Oy. About the development environment, see the demo report p. 15 (Rouhuvirta et al., Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Codacmos 2004. Project IST-2001-38636).

LAJI	MAARA	S_RIV1
1	28894.69	Palkka päätoimesta
2	7215.27	Palkasta
4	40.44	Yhtiöveron hyvitys
7	235.47	Puun myyntitulosta
40	700	Jäsenmaksut
86	99	Osingot 00-03
87	40.44	Yhtiöver hyv 29/71
132	27048.32	VEROVUODEN AT W
134	25015.69	VEROVUODEN AT KV
136	1009.13	Eläketulot yht
141	1316.76	VEROVUODEN POT
190	2790.63	Tulovero väh jälk
420	1177.32	Metsät puhd pot

Search from XML: PALKKA Match finder: 44444

Figure 5. Taxpayer's tax record selected from the relational database (user interface screen)

PALKKA

description

- title
 - Mikä on palkkaa
- paragraph
 - keyword
 - Palkka
 - Palkan käsite on määritelty EPL 13 §
 - list
 - listitem
 - listpara
 - 1) kaikenlaatuista palkkaa
 - listpara
 - 2) kokouspalkkiota, henkilökohtaista luento- ja esitelmäpalkkiota, hallintoelimen jäsenyydestä saatua palkkiota, toimitusjohtajan palkkiota, avoimen yhtiön ja

PALKKA

Mikä on palkkaa

Palkan käsite on määritelty EPL 13 §:ssä (1118/96). Sen mukaan palkalla tarkoitetaan:

- 1) kaikenlaatuista palkkaa, palkkiota, etuutta ja korvausta, joka saadaan työ- tai virkasuhteessa;
- 2) kokouspalkkiota, henkilökohtaista luento- ja esitelmäpalkkiota, hallintoelimen jäsenyydestä saatua palkkiota, toimitusjohtajan palkkiota, avoimen yhtiön ja

PALKKA Search

Figure 6. Metadata of register data (user interface screen)

27. Metadata is presented (see Figure 6.) as a structure and an XML document. The structure follows the concept hierarchy and it can be browsed along the hierarchy in both directions. By browsing the structure it is possible to go deeper from a more general concept to the concepts it contains and their

metadata. Thus it is possible to move by browsing from a more general concept, such as earned income, to pay and its metadata and from wages to the concept of shunting and its metadata.

28. The hierarchy is made similarly both for taxable items and deductions made from them. In addition, the structure of the metadata enables in different hierarchy levels descriptions of the legal cases concerning the section in question and references to the legislation regulating it.

VIII. STRUCTURED METADATA – STATISTICS

29. Using of structured register metadata as part of statistical metadata is illustrated here by an example from income distribution statistics. The data for income distribution statistics are mostly collected from administrative registers, one of the key data sources being the personal taxation register. The way in which the register data is used in the compilation of the income distribution statistics is defined in the income formation rules. They indicate which register item is used for forming which income concept. In its present scope the metadata of the income distribution statistics describes the definition of the concept and its possible source.

30. The key concept of the income distribution statistics is the concept of disposable income. When the contentual description information of the concept is arranged according to the CoSSI metadata specification into structured metadata, the concept's metadata can be reduced to the figure below (see Figure 7.).

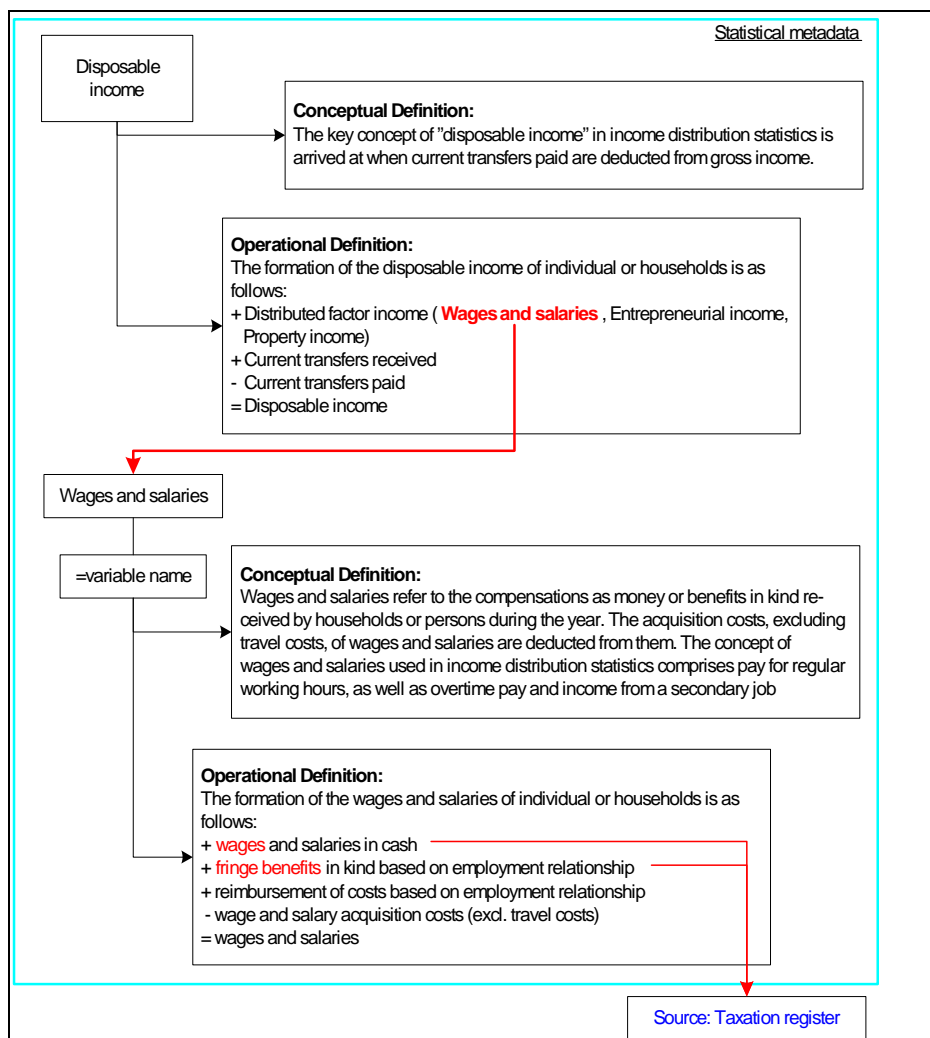


Figure 7. Statistical metadata of the income concept

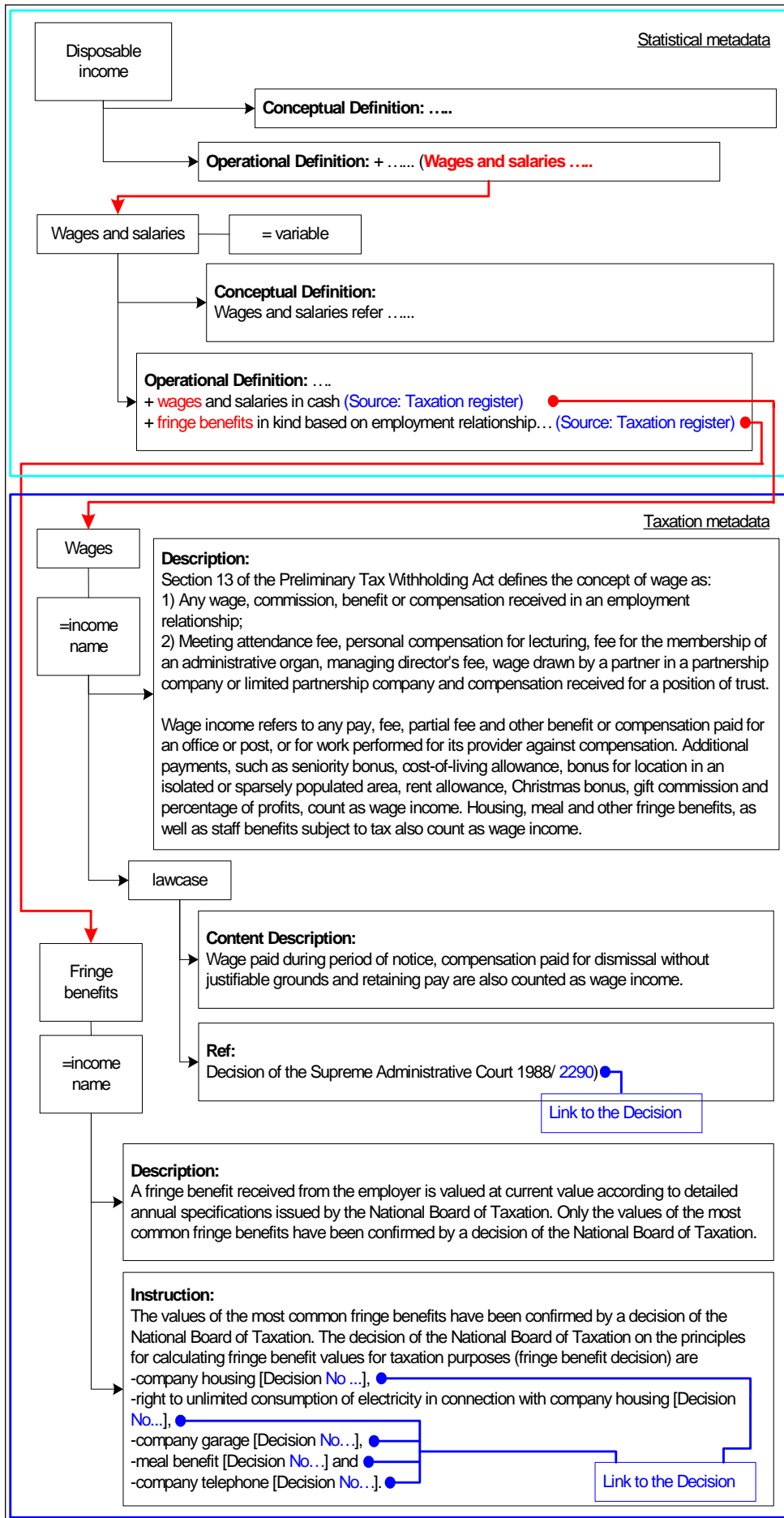


Figure 8. Extended statistical metadata of the income concept

31. The main factor collecting and transmitting the information in the structured statistical metadata is variable. When for register data there is available structured metadata as that produced in the development environment, the register metadata can be connected by quite simple procedures as part of statistical metadata when using register data. The resulting metadata is illustrated in Figure 8, which shows in a simplified manner the metadata of the disposable income concept extended by combining and corresponding to the concept formation of the income distribution statistics.

32. Thus arranged, metadata makes the processing rules and procedures followed in taxation available to statistics producers editing the register data and the tax authorities' direct information influencing the interpretation can be used in statistics production during editing as well. The created metadata structure also enables linking of original sources to the metadata, whereby this material can also be used by statistics producers. These links to the legislation database or guidelines, for example, directed to both taxpayers and taxation officials for reporting and revising tax information can be opened directly and their content can be viewed by statistics producers.

33. Combination of the metadata can be done only when the metadata are structured as presented above. At the same time it will become possible for statistics production to automate the handling process of metadata and to exploit structured technologies in other ways as well.

IX. TOWARDS METADATA-DRIVEN STATISTICAL PRODUCTION

34. When the register metadata can be integrated as shown above into statistical metadata in the production process of statistics, it opens the possibility to develop a genuinely metadata-driven

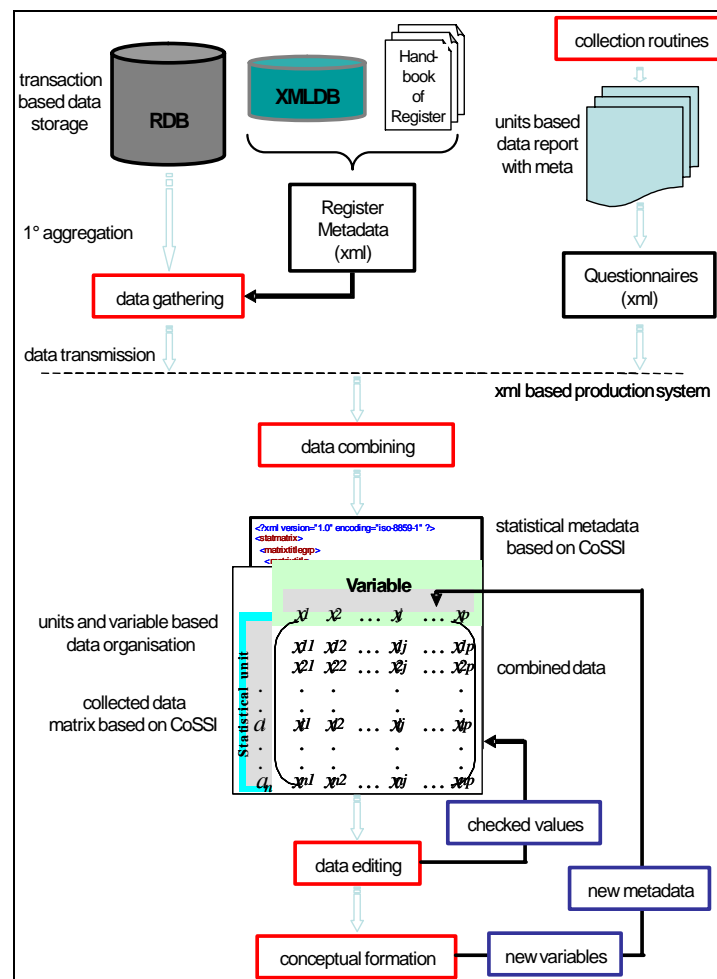


Figure 9. Compilation, combination and editing of collected information in the CoSSI reference frame

production process. In a genuinely metadata-driven production process rich metadata is present and available in all production stages, including editing, and metadata accumulates as the process advances without losing old metadata. As a model the process can be described as in Figure 9.

35. In the model the selection of register data and collection of primary data both produce structured metadata. Metadata is transferred to the editing stage of statistical production where the data are checked and edited in many ways. During editing new conceptual variables and metadata describing them are formed to the data. Thus we end up with final statistical data, which also contain statistical metadata. When tables and other outputs are produced from the data by tabulation, the metadata necessary for the interpretation of statistical figures can be included fully in statistical tables, similarly as in other outputs.

References

Maler, E. & El Andaloussi, J. 1996. Developing SGML DTDs. From text to model to markup. Upper Saddle River (NJ), USA: Prentice Hall.

Rouhuvirta H., An alternative approach to metadata – CoSSI and modelling of metadata, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636. Available on the web at: http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf

Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.9, Statistics Finland 2003. Available on the web at: http://www.stat.fi/org/tut/dthemes/drafts/cossi_definition_descriptions_v_09_2003.pdf

Rouhuvirta, H. and Lehtinen, H., Taxmeta DTD. Codacmos 2004. Project IST-2001-38636. Available on the web at: http://www.stat.fi/org/tut/dthemes/drafts/taxmeta_dtd_v_01.txt

Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S., Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Codacmos 2004. Project IST-2001-38636. Available on the web at: http://www.stat.fi/org/tut/dthemes/papers/demoreport_on_taxation_metadata_codacmos_2004.pdf